Character and Document Research in the Open Mind Initiative

David G. Stork
Ricoh Silicon Valley
2882 Sand Hill Road #115
Menlo Park, CA 94025-7022
stork@rsv.ricoh.com

Abstract

We describe the Open Mind Initiative, a framework for large-scale collaborative efforts in building components of "intelligent" systems that address common-sense reasoning, document and language understanding, speech and character recognition, and so on. Based on the Open Source methodology, the Open Mind Initiative allows domain specialists to contribute algorithms, tool developers to provide software infrastructure and tools, and non-specialist "e-citizens" to contribute training data and information to large databases. An important challenge is to make it easy and rewarding for e-citizens to provide such information. This paper illustrates the Initiative through several demonstration projects of modest scale, including some related to character and document problems, and identifies general challenges and opportunities.

1 Introduction

"Intelligent" machines — ones that can understand speech, read handwriting, recognize objects and actions from images or video clips, summarize stories, reason about the world, engage in conversation, play world-class chess or go — have been a dream and research goal for the last half-century. Efforts to build such systems have relied on both theory and data [11, 27], and there has been slow incremental improvement in a number of areas. We argue that several disciplines would profit from very large amounts of data and an open framework for experimentation and collaboration, and point to a new methodology — The Open Mind Initiative — to this end.

The paper is organized as follows: In Sect. 2 we contend that current models are adequate or nearly adequate for many important pattern recognition and knowledge engineering problems, and that it is the lack of training data and an open framework for systems engineering and integration that retards progress. Even in domains where the models may not yet be adequate, large amounts of knowledge and training data are needed to decide between competing models and thus to accelerate progress. In Sect. 3 we show that such limitations may be overcome in part by a new software methodology, Open Source, and the infrastructure of the web. This leads to a description of The Open Mind Initiative in Sect. 4. An important distinction with Open Source is the reliance of Open Mind on in-

frastructure and tools and large numbers of relatively untutored contributors. The Initiative is illustrated in Sect. 5 with several proposed projects, including two from character and document related areas. The paper concludes in Sect. 6 with some general remarks and important challenges.

2 Models, tools and data

All pattern recognition and intelligent systems rely on theory and models as well as a knowledge base or training data; such systems (particularly fielded commercial ones) also rely on a great deal of software engineering.

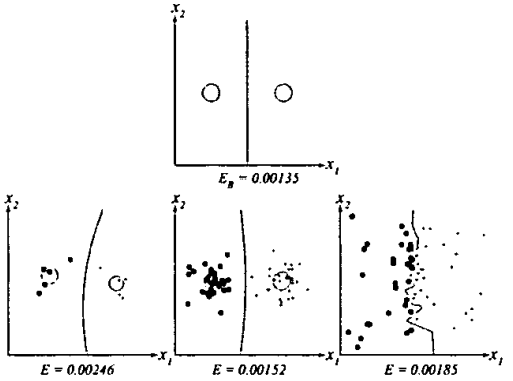

Figure 1: General, non-specific models trained with a large amount of data can outperform better models trained with a small amount of data (see text).

2.1 Models

In very broad terms, recent work in many areas of pattern recognition and artificial intelligence has relied increasingly upon fairly general models, such as powerful statistical ones, trained with a great deal of data. The fundamental theoretical underpinnings of domain-independent pattern recognition — maximum-likelihood and Bayesian techniques, function estimation, and so on — are highly developed and rigorous. While there will continue to be effort and progress, the foundations as currently understood are sufficient for developing successful pattern classifiers in many domains.

The adequacy of even very simple models is illustrated in optical character recognition, where accurate recognizers can use simple models (decision trees, neural networks, ...) trained with millions of characters. These outperform recognizers based on sophisticated models trained with less data. As Ho and Baird conclude, "quality of training data is the dominant factor affecting accuracy" and "as long as the training data are representative and sufficiently many, a wide range of classifier technologies can be trained to equally high accuracies" [15], a point stressed elsewhere independently [28, 13]. The key lesson from decades of work in isolated character recognition is this: Even a weak (but general) model can yield excellent results if it is trained with sufficient data.

This fact is illustrated schematically in Fig. 1. The top figure shows the ground truth for a two-dimensional, two-category classification task with two equal-variance circular Gaussian prior distributions. The Bayes error, E_B, is shown. The figure at the left shows excellent models, circular Gaussians, whose means and variances were trained by maximum likelihood with a small number of points; the resulting decision boundary and test error are shown as well. The center figure shows a less-specific model, where each Gaussian can have *arbitrary* covariance matrix; these Gaussian models were trained with a large amount of data, however. The final classifier has a test error lower than that of the better model at the left. The figure at the right shows the decision boundary of a nearest-neighbor classifier trained by learning with queries, where

the points presented to users were generated as mid-way between two points, one chosen randomly from each category. Because a large number of user-classified points are used as training data, even this simple classifier outperforms the better model at the left. In short, a very simple nearest-neighbor classifier trained interactively and a weak general model trained with a large amount of data outperform an excellent model trained with a small amount of data.

This need for large training sets is a lesson that recurs in a number of domains, from acoustic speech recognition [19] and speechreading [33] to face recognition [23], gesture recognition [24], natural language processing [10], speech production [35] and others. Moreover, for areas where we may not yet have adequate models, we know how to broaden and improve classes of models — to include more degrees of freedom to account for sources of variation, to set parameters, and so on — given enough data. Again, in broad terms, it appears that after decades of work we now have (or know how to create) adequate models in some disciplines and that progress is retarded by the lack of adequately large knowledge bases and training data.

2.2 Infrastructure and tools

Another key component in building such systems involves software tools. There are many commercial tools for developing speech and natural language systems which allow developers to explore model parameters easily, specify grammars, lexicons, and so on. Some of these tools are provided free of charge by companies in order to promote the sale and use of their hardware, such as speech chips [34, 16]. An important lesson here is that non-specialist developers can create useful systems, given sufficiently good tools.

2.3 Training data

In some domains, the need for training data is addressed with synthetic data, in which raw sensed patterns are transformed in model-based ways to yield surrogate ones. For example, in optical character recognition synthetic data is constructed by automatically rotating, warping, line thinning or thickening, and adding pixel noise to sensed patterns. Synthetic data is attractive as it is far simpler to obtain than an "equivalent" number of raw sensed patterns, though some controversy remains as to their relative merits. One great benefit of synthetic data is in learning with queries. Here synthetic patterns can be generated "on demand" in informative regions of feature space, i.e., near decision boundaries.

The appreciation of the need for large knowledge bases and training data has led to the construction of publicly available databases. The National Institutes of Standards and Technology (NIST), the Linguistics Data Consortium (LDC), The Center for Excellence in Document Analysis and Recognition (CEDAR), Information Science Research Institute (ISRI), the University of Washington CD-ROM and other sites have compiled large databases of training data related to language and documents. One of the largest comes from the Macrophone project, compiled by Texas Instruments, a collection of roughly 200,000 utterances of free telephone speech by non-specialists, constrained by topic [6]. While these and other public databases have been vital to continued improvements in recognizers, some of the best systems are trained with additional data, usually proprietary [7]. This need for data is not restricted to training data in the classical pattern recognition framework, but includes more general information, such as diseases and their medical symptoms [31], common sense facts about the world [21], synonyms and relations, etc. [12].

3 Open Source Model

Recent technical, social and economic developments suggest a new approach to augmenting knowledge databases and to building intelligent systems. A decade or two ago, none but a few eccentric visionaries could imagine that large-scale highly reliable software could be developed outside of corporate, government or university research labs. Nevertheless, the Free Software Foundation and now particularly the Open Source movement — which promotes software reliability and quality by supporting independent peer review and rapid evolution of source code distributed for free — have blossomed from curiosities to significant trends. These trends have already influenced the world software market (particularly in operating systems and internet software) and show no sign of abating. Leading Open Source software includes: *Linux*, a Unix-like operating system (with nearly 100 million lines of code and over 10 million installed seats, built by roughly 100,000 contributors [26]); the *Mozilla* version of the Netscape Navigator Web browser (several hundred thousand lines); *SendMail*, a utility on virtually every Unix machine; and *Apache*, which runs over half of the world's web servers. A particularly relevant example here is the Open Source web indexing produced by *Newhoo* (purchased by Netscape/AOL). In the *Newhoo* approach, numerous non-specialist contributors propose keyword and index information about web pages; this information is reviewed by volunteer referee/editors (currently 4600), collected and made available to all. The software infrastructure and tools themselves were developed through the Open Source method. While proprietary software generally improves logarithmically with time, Open Source generally improves exponentially.

4 Open Mind Initiative

In light of these developments we can approach the creation of intelligent systems based in part on the Open Source model — the Open Mind Initiative. It is perhaps simplest to consider its three component functions — provided by domain experts, infrastructure and tool developers, and e-citizens — corresponding roughly to the three headings in Sect. 2. Domain experts contribute libraries of fundamental algorithms; tool developers contribute and refine the enabling software; e-citizens contribute information and training data. Users with an interest and expertise in a particular domain such as speech, vision, language, common sense, and so forth could serve as reviewers or moderators. All this is possible given the infrastructure of the internet and world wide web.

As in Open Source, the Open Mind Initiative would be only loosely structured, and there is no clear notion of hierarchy. Individuals might participate in several stages at different times and each project will require a different level of effort on the various components. There would be a number of component projects, with varying amounts of interactions.

4.1 Domain experts

Experts in a specific area such as optical character recognition or speech recognition will submit documented libraries of fundamental algorithms, and possibly representative training sets, all in Open Source and freely available to all. Much of this work has already been published in refereed journals. This approach extends the trend in academic publishing in which algorithms and data are published in electronic form on the web.

4.2 Tool developers

Key components to the Initiative are provided by the infrastructure and tool developers; these components differ somewhat from those in traditional software engineering and research. Aside from user interfaces, format conversion, and so on, software will have to detect significant errors or "outliers" in contributed data for review and possible elimination. The distinct challenge, however, is to make it easy for e-citizens to contribute data, and here new forms of infrastructure will have to be developed. Consider *Animals*, an interactive children's program from the late-1970s and early 1980s for classifying animals [30]. The child thinks of an animal, and the program tries to guess it from the child's responses to a series of questions, such as ABLE TO FLY (YES OR NO)?, TWO LEGGED OR FOUR LEGGED?, and so forth. After the final question, the program makes a guess, for instance CHICKEN. If the guess is wrong (the child was thinking instead of TURKEY), the child must provide a new question that distinguishes her animal from the one guessed by the program, e.g., RED COMB ON HEAD?. This new query is automatically incorporated into the program. After a number of children have played this guessing game, *Animals* has learned a simple tree-based classifier for animals. We have written a Java-based program, *Open Mind Animals*, in order to develop tools and infrastructure useful in the Open Mind Initiative [20].

Similarly, the *Answer Garden* project has demonstrated that knowledge bases for help desk environments can be grown "organically" and semi-automatically [1, 2]. Perhaps new and more sophisticated games, inspired by *Animals* and *Answer Garden* could be part of the Open Mind Initiative, and would encourage large numbers of users to contribute with minimal burden. This basic approach is already exploited in *DirectHit*, which refines web indexes based on actual user search sequences. We call this approach "unconscious knowledge capture." Here expertise in design, advertising and marketing would be needed to complement more traditional software skills.

In such a relatively unstructured massive collaborative software project many technical problems must be considered — everything from low-quality data to outright hostile attacks. A number of simple heuristics in data "truthing" could reduce the possibility of poor data. For instance, any query from the Open Mind system could be presented to three independent, randomly chosen users, and the reply accepted if all three agree. Likewise, there are domain-dependent algorithms for automatically identifying "outliers" — responses that differ drastically from the current consensus. Such outliers could be brought automatically to the attention of a moderator/referee for review. There are many techniques from experimental psychology for insuring the quality of the data, such as the insertion of a "catch trial" which has only one plausible answer; an incorrect answer on such a catch trial belies an unreliable contributor and invalidates his or her recent submissions.

4.3 E-citizens

The biggest difference between traditional Open Source and the Open Mind Initiative is the need for data provided by e-citizens. Such contributions have been made in other fields, and given the right circumstances should occur in the Open Mind Initiative too. For instance, amateur astronomers discover comets and contribute measurements of variable stars; over 13,000 amateur ornithologists participate in bird counts annually; amateurs scoured innumerable books for words to contribute to the massive Oxford English Dictionary project [25]; amateur paleontologists discover fossils

and promising excavation sites; over half of the mollusks in public museums come from amateurs [22]; and so forth. When NASA's High Resolution Microwave Survey project (which included the Search for Extraterrestrial Intelligence) was canceled, the non-profit SETI Institute was founded and thousands of individuals donated time on their personal computers to search snippets of radio telescope recordings for tell-tale signs of an intelligent life [32]. E-citizens might be motivated to make a meaningful contribution to a research project such as the Open Mind Initiative, perhaps through schools. Even if a tiny percentage of people with web access provide a small amount of information, the training sets can grow by three or four orders of magnitude. For these people there is a need to provide a clear, realistic, yet inspiring vision of the project and its value to science and society.

While contributing e-citizens should be permitted to remain anonymous, others might be glad to see their name listed in the Open Mind Initiative website, perhaps ranked by the amount of their submitted information was accepted into the system. It was found, for example, that in a collaborative effort for developing documentation for field service, service professionals actually preferred community recognition and improved reputation (through posting of their name) over monetary incentives, which were viewed as corrupting the process [5]. A somewhat less lofty but probably effective inducement could be provided by sales or service companies, who would give discounts or benefits (e.g., frequent flyer miles) proportional to the amount of information an e-citizen contributes. In this way, the participating companies would attract new customers. Likewise a lottery could be instituted, in which each submission of data gives its contributor a small chance of winning a large prize.

Regardless of what motivates someone to contribute, all will want to see how the system progresses, be it by some raw measure of the total amount of information contributed, or in the quantitative performance of one of its components such as recognition accuracy. Other, qualitative indications of progress would be useful too. Just as parents delight in watching the cognitive development of their child, so too would contributors be excited to see an Open Mind common-sense reasoning system develop its "understanding" of the world.

5 Sample projects

Here we sketch three proposals for projects that seem well suited to the Open Mind Initiative: handwritten character recognition, handwritten word recognition and domain knowledge organization. A fourth is already in development and will be reported elsewhere [20].

5.1 Isolated handwritten character recognition

The task is the recognition of isolated handwritten characters, for instance displayed as 8×8 graylevel pixel images. The classifier could be very simple indeed, such as a basic decision tree [8] or a neural network; the effort will center on infrastructure, tools and the social and organizational problems. Suppose that while your web browser boots up, it displays an image of a handwritten character, along with two buttons each labeled by a category, e.g., "4" and "9." You click on the button corresponding to your perception of the image, and your reply is sent automatically to an Open Mind repository, a tiny contribution toward improved character recognizers. Training could be efficient since the system would present to e-citizens only ambiguous patterns (i.e., most informative), a technique known as learning with queries [3, 4]. This technique

often provides a distinct advantage over traditional i.i.d. sampling in that data can be provided near decision boundaries (cf. right figure in Fig. 1). Other such informative patterns could be generated by transforming raw patterns (line thinning, rotation, skew, etc.). There should be an interface to enable very active contributors to quickly and easily provide a large number of responses. With millions of such replies (possibly in different contexts), the system's recognizer would become accurate indeed.

One method for "truthing" — detecting and reducing poor quality data, eliminating grossly misshapen patterns, and so forth — would be to present any individual character image to three independent, randomly chosen contributors, and accept their result only if all three agree. Other, more sophisticated and automatic methods can be used to detect outliers for rejection [14], avoid statistical bias, and so on.

Contributors might include a portion of the members of laboratories working in the general field, their home institutions and related communities (≈100), students in pattern recognition or other related courses worldwide (≈200), members of the *Linux* and Open Source community (≈1000), and interested e-citizens (≈200) made aware through short notices on discussion or mailing lists, links from personal or lab home pages or broad public announcements. As mentioned above, contributors may be more motivated to contribute if they can see the total amount of data submitted as well as comparisons with a state-of-the-art classifiers as the Open Mind classifier improves over time.

5.2 Handwritten word recognition

A closely related problem would be recognition of handwritten words. Here the data would come from a large database of scanned handwritten documents automatically segmented by current algorithms. The e-citizen contributors are either given a choice of candidate words from a classifier [18] (presented as clickable buttons) or are asked to type in their transcription. Here too, only the most informative or ambiguous word images would be presented to the e-citizens, and nonsense or poorly segmented words would be marked for elimination by e-citizens.

5.3 Knowledge engineering

A particularly interesting and instructive project could be based on CYC, a pioneering attempt to capture common-sense knowledge. For over a decade, roughly a dozen CYC knowledge engineers have been entering more than 400,000 assertions (or "rules") designed to capture a significant portion of our consensus knowledge about the world [21]. For instance, CYC knows that a mother is older than her son, that clouds are usually outdoors, and that a cup filled with wine will be rightside-up, not upside-down. As part of its training, CYC "reasons" about its information, searching for inconsistencies or ambiguities, which are then presented as questions to be answered by the knowledge engineers. These answers, along with further rules, are recompiled and the cycle repeats. Its designers believe that once CYC attains a large amount of common-sense information, it can then continue to learn by "reading" digital encyclopedias or books.

In an Open Mind project inspired by CYC, if the questions were posed in a clear way to even a fraction of the web population, large amounts of common sense knowledge could be entered rapidly. Done carefully, this would provide confirming or disconfirming evidence for the CYC model of knowledge representation. Rather than begin with a project so ambitious as general common sense, however, it

would be more productive to limit the domain of discourse, for instance to computers and software. Thus the system would learn "common sense" knowledge, such as the need for programs to be compiled or interpreted, that early versions of code are often buggy, that a mouse is a peripheral, and so on. The domain experts would set the data structures and conflict resolution algorithms so that the ontology of the domain can be captured. We can expect that many computer-savvy contributors would be motivated to provide such information and would be early users of any final system. A great challenge for the infrastructure providers would be to develop an interface, or possibly cast the data acquisition as a game.

5.4 Future projects

There are a number of projects that could take advantage of an Open Mind Initiative, varying in the scope, difficulty, scientific or practical usefulness, and so on. Since many OCR systems based solely on pixel images asymptote at a fairly high but not perfect accuracy (e.g., 95%), there is a clear need for algorithms for subsequent stages. Grammar, syntax, context and topic identification algorithms, developed through Open Mind, would be valuable here. While of modest scientific value, an Open Mind effort to develop an entry to the Loebner Prize ("Turing test") for the most "humanlike" dialog system would present an interesting challenge to interface and database design. For instance, consider a Dungeons and Dragons game in which the goal is to navigate through a castle to the "human" king while avoiding the "robot" king. In each room, you are presented with two short paragraphs, each "written" by one of the kings (i.e., by natural language text generation algorithms with different parameters). You then click on the paragraph that seems most "humanlike." Each choice is used to refine the computer models of word frequency, sentence and phrase structure, and so on, to improve the model of human-like language.

Computer chess systems rely on massive parallel search guided by numerical scores of each board position. Some systems, such as IBM's *Deep Blue*, stress search over sophisticated board scoring [17]. An Open Mind chess project would allow interested amateurs to rate a large number of board positions and thereby improve the system's beam search. This might lead to a more "human" style of chess play. The reliability of a contributor might be based on public chess ratings or performance on an on-screen test. Because the branching factor for search the Chinese/Japanese board game go is so high, go is unlikely to succumb to brute force approaches. Perhaps *Open Mind Go*, relying on a large number of board scores contributed by go players, would provide the first serious challenge to human go masters.

Building upon lessons learned from the knowledge engineering project described in Sect. 5.3, we can expect analogous projects in domains such as finance, office systems, sports, and so forth. One of the greatest benefits of the open nature of the Initiative is that it would facilitate the integration of different projects. Thus, speech recognition could be integrated with natural language and common sense databases for improved human-machine interfaces, smarter web searching, and so on.

5.5 Business and legal issues

Just as economic and commercial cases can be made for Open Source, so too can they be made for the Open Mind Initiative. Commercial firms could supply the resulting software directly or provide customization and support, as does Red Hat Software, Inc., for *Linux*. Hardware manufacturers could build devices that run Open Mind software, thereby reduc-

ing their software costs. Perhaps the most important benefit would be to open market niches (e.g., for common sense software) that few if any individual commercial software companies could provide economically.

The legal matters would have to be considered in order to avoid intellectual property disputes that have plagued big science projects such as the Human Genome Project. Perhaps since the data was freely contributed, it should be free for use, even in commercial settings. Thus a licensing agreement modeled on that for Berkeley FreeBSD might be appropriate.

6 Related work

There are several important differences between Open Source/Free Software and the proposed Open Mind Initiative, as shown in Table 1.

Open Source	Open Mind
no e-citizens	e-citizens crucial
expert knowledge	informal knowledge
machine learning irrelevant	machine learning essential
web optional	web essential
most work is directly on the final software	most work is *not* on the final software
hacker/programmer culture ($\approx 10^5$)	e-citizen/business culture ($\approx 10^8$)
separate functions contributed (e.g., *Linux* device drivers)	single function goal (e.g., OCR recognition)

Table 1: Comparison of Open Source and Open Mind approaches.

In traditional Open Source each contributor typically provides code which addresses a different functionality, for instance device drivers and routines for handling specific file formats in *Linux*. In Open Mind, however, the goal is generally a *single* function, such as accurate speech recognition. This distinction implies a somewhat different role for oversight in the two cases. In Open Mind, moderators should monitor the improvement on the single goal, rather than just compatibility of a proposed addition in Open Source.

Table 2 compares traditional data mining over the web with the Open Mind approach. One significant difference here is that the needed data might be unavailable in data mining application, for instance OCR data is not available by data mining the web itself.

Data Mining	Open Mind
needed data might be unavailable	data tailored to the project (e.g., OCR)
no queries	interactive queries
ambiguities may never be resolved	ambiguities may be resolved by queries
relatively fixed amount of data	new data encouraged (feedback loop)
slow learning	faster learning
little or no e-citizen support	e-citizen support essential

Table 2: Comparison of Data Mining and Open Mind approaches.

7 Big science

Physics has had its atom smashers, Aeronautics and Astronautics its space missions, Microbiology its Human Genome Project. It is time for cognitive science, computer science, pattern recognition, artificial intelligence and related fields to have our big science — one based on a Open Source model in which all citizens can contribute a bit of their perception and understanding of the world. Open Mind would cost a fraction of other big science projects yet be supremely useful. The only

"big" computer science project of this general nature (outside the military and national security area [9]) is the Digital Libraries Initiative [29], which addresses explicit formal knowledge. In contrast, the Open Mind Initiative would capture the implicit informal knowledge that all people have.

Given the conjunction of several forces — the need for applications such as natural human-machine interfaces and improved web searching, the existence of good theory, good infrastructure, demonstrated success of the Open Source methodology — the time seems right for projects in the Open Mind framework. Lessons learned from the projects described in Sect. 5 will be invaluable for further progress.

A plausible model of the human brain is a collection of modules, each evolved, trained and selected to address a small set of tasks. In a loose, metaphorical way, the Open Mind Initiative would build a "brain" in an analogous way.

Acknowledgements

The author would like to acknowledge discussions and correspondence with Marko Balabanovic, Mindy Bokser, Ron Cole, Jonathan Hull, David Israel, Yann Le Cun, George Nagy, Eric Raymond, and Richard Schwartz. The views expressed here are those of the author and not necessarily of those just listed.

References

[1] Mark S. Ackerman and Thomas W. Malone. Answer Garden: A tool for growing organizational memory. In *Proceedings of the Conference on Office Automation Systems*, Filtering, Querying, and Navigating, pages 31–39, New York, NY, 1990. ACM Press.

[2] Mark S. Ackerman and David W. McDonald. Answer Garden 2: Merging organizational memory with collaborative help. In *Proceedings of the ACM 1996 Conference on Computer Supported Work*, pages 16–20, New York, NY, 1996. ACM Press.

[3] Dana Angluin. Queries and concept learning. *Machine Learning*, 2(4):319–342, 1988.

[4] Dana C. Angluin. Learning with queries. In Eric B. Baum, editor, *Computational Learning and Cognition*, pages 1–28, Philadelphia, PA, 1993. SIAM.

[5] David G. Bell, Daniel G. Bobrow, Olivier Raiman, and Mark H. Shirley. Dynamic documents and situated processes: Building on local knowledge in field service. In Toshiro Wakayama, Srikanth Kannapan, Chan Meng Khoong, Shamkant Navathe, and JoAnne Yates, editors, *Information and Process Integration in Enterprises: Rethinking Documents*, pages 261–276, Boston, MA, 1998. Kluwer Academic Publishers.

[6] Jared Bernstein, Kelsey Taussig, and Jack Godfrey. Macrophone: An American English telephone speech corpus for the Polyphone project. In *Proceedings of the International Conference on Automatic Speech and Signal Processing (ICASSP94)*, volume I, pages 81–84, Adelaide, Austrailia, 1994.

[7] Mindy Bokser, 1999. Personal communication (Caere Corporation).

[8] Leo Breiman, Jerome H. Friedman, Richard A. Olshen, and Charles J. Stone. *Classification and regression trees*. Chapman & Hall, New York, NY, 1993.

[9] Paul Cohen, Robert Schrag, Eric Jones, Adam Pease, Albert Lin, Barbara Starr,

David Gunning, and Murray Burke. The DARPA high-performance knowledge bases project. *AI Magazine*, 19(4):25–49, 1998.

[10] Walter Daelemans, Antal van den Bosch, Jakub Zavrel, Jorn Veenstra, Sabine Buchholz, and Bertjan Busser. Rapid development of NLP modules with memory-based learning. In Roberto Basili and Maria Theresa Pazienza, editors, *ECML98 TANLPS Workshop Notes*, pages 1–17, Technische Universität Chemnitz, 1998.

[11] Richard O. Duda, Peter E. Hart, and David G. Stork. *Pattern Classification.* Wiley, New York, NY, second edition, 2000.

[12] Christiane Fellbaum, editor. *WordNet: An Electronic Lexical Database.* MIT Press, Cambridge, MA, 1998.

[13] Isabelle Guyon, John Makhoul, Richard Schwartz, and Vladimir Vapnik. What size test set gives good error rate estimates? *IEEE Pattern Analysis and Machine Intelligence*, PAMI-20(1):52–64, 1998.

[14] Thien M. Ha. Efficient detection of abnormalities in large OCR databases. In *Proceedings of the International Conference on Document Analysis and Recognition (ICDAR97)*, volume 2, pages 1006–1010, Los Alamitos, CA, 1997. IEEE Press.

[15] Tin Kam Ho and Henry S. Baird. Large-scale simulation studies in pattern recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-19(10):1067–1079, 1997.

[16] Jerry R. Hobbs, Douglas Appelt, John Bear, and David Israel. FASTUS: A system for extracting information from text. In *Proceedings of the ARPA Human Language Technology Workshop '93*, pages 133–137, Princeton, NJ, 1994. Distributed as *Human Language Technology* by San Mateo, CA: Morgan Kaufmann Publishers.

[17] Feng-hsiung Hsu, Thomas Anantharaman, Murray Campbell, and Andreas Nowatzyk. A grandmaster-level-chess machine. *Scientific American*, 263(4):44–50, 1990.

[18] Jonathan H. Hull, Tin Kam Ho, John Favata, Venu Govindaraju, and Sargur N. Srihari. Combination of segmentation-based and wholistic handwritten word recognition algorithms. In Sebastiano Impedovo and Jean-Claude Simon, editors, *From Pixels to Features III: Frontiers in Handwriting Recognition*, pages 261–272, New York, NY, 1992. Elsevier-North Holland.

[19] Frederick Jelinek. *Statistical Methods for Speech Recognition.* MIT Press, Cambridge, MA, 1998.

[20] Chuck Lam and David G. Stork. Open Mind Animals, 1999. Java program and documentation.

[21] Doug B. Lenat. CYC: A large-scale investment in knowledge infrastructure. *Communications of the ACM*, 38(11):33–38, 1995.

[22] James H. McLean, 1999. Personal communication from the Los Angeles County Museum of Natural History.

[23] Baback Moghaddam, Tony Jebara, and Alex Pentland. Bayesian modeling of facial similarity. In *Advances in Neural Information Processing Systems*, volume 10. MIT Press, Cambridge, MA, 1998.

[24] Baback Moghaddam and Alex Pentland. Maximum-likelihood detection of faces and hands. In *Proceedings of the International Workshop on Automatic Face- and Gesture-Recognition*, pages 122–128, Zurich, Switzerland, 1995.

[25] K. M. Elisabeth Murray. *Caught in the Web of Words: James A. Murray and the Oxford English Dictionary.* Yale University Press, New Haven, CT, reprint edition, 1995.

[26] Eric S. Raymond, 1999. Personal communication from `OpenSource.org`.

[27] Stuart Russell and Peter Norvig. *Artificial Intelligence: A Modern Approach.* Prentice Hall Series in Artificial Intelligence. Prentice Hall, Englewood Cliffs, NJ, 1995.

[28] Michael Sabourin, Amar Mitiche, Danny Thomas, and George Nagy. Hand-printed digit recognition using nearest neighbors classifiers. In *Proceedings of the Second Annual Symposium on Document Analysis and Information Retrieval*, pages 397–409, Las Vegas, NV, 1993.

[29] Bruce Schatz and Hsinchun Chen. Building large-scale digital libraries. *IEEE Computer*, 29(5):22–26, 1996.

[30] Stuart C. Shapiro, January 1982. Programming Project 1: Animal Program (20 Questions), "Introduction to Artificial Intelligence," Department of Computer Science, State University of New York at Buffalo.

[31] Edward H. Shortliffe. *Computer-Based Medical Consultations: MYCIN.* Elsevier/North-Holland, New York, NY, 1976.

[32] Seth Shostak. *Sharing the Universe: Perspectives on Extraterrestrial Life.* Berkeley Hills, Berkeley, CA, 1998.

[33] David G. Stork and Marcus E. Hennecke, editors. *Speechreading by Humans and Machines: Models, Systems, and Applications.* NATO Advanced Studies Institute. Springer, New York, NY, 1996.

[34] Stephen Sutton, Ron A. Cole, Jacques de Villiers, Johan Schalkwyk, Pieter Vermeulen, Michael Macon, Yonghon Yan, Ed Kaiser, Brian Rundle, Kal Shobaki, Peter Hosom, Alex Kain, Johan Wouters, Dominic Massaro, and Michael Cohen. Universal speech tools: The CSLU Toolkit. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP98)*, pages 3211–3224, Sydney, Australia, November 1998.

[35] Antal van den Bosch and Walter Daelemans. Data-oriented methods for grapheme-to-phoneme conversion. In *Proceedings of the Sixth Conference of the European Chapter of the ACL*, pages 45–53. ACL, 1993.

Character and document research in the Open Mind Initiative

Stork, D.G.
Ricoh Silicon Valley, Menlo Park, CA;

This paper appears in: **Document Analysis and Recognition, 1999. ICDAR '99. Proceedings of the Fifth International Conference on**
Meeting Date: 09/20/1999 -09/22/1999
Publication Date: 20-22 Sep 1999
Location: Bangalore , India
On page(s): 1-12
References Cited: 35
IEEE Catalog Number: PR00318
Number of Pages: xxiv+821
INSPEC Accession Number: 6352780

Abstract:
We describe the Open Mind Initiative, a framework for large scale collaborative efforts in building components of "intelligent" systems that address common sense reasoning, document and language understanding, speech and character recognition, and so on. Based on the Open Source methodology, the Open Mind Initiative allows domain specialists to contribute algorithms, tool developers to provide software infrastructure and tools, and non specialist "e-citizens" to contribute training data and information to large databases. An important challenge is to make it easy and rewarding for e-citizens to provide such information. The paper illustrates the initiative through several demonstration projects of modest scale, including some related to character and document problems, and identifies general challenges and opportunities

Index Terms:
character recognition common-sense reasoning database management systems document handling knowledge based systems speech recognition Open Mind Initiative Open Source methodology character recognition common sense reasoning document problems document research domain specialists e-citizens intelligent systems language understanding large databases large scale collaborative efforts software infrastructure tool developers training data